Class 17: Rademacher Averages and Symmetrization

Alexander Rakhlin

This class is based largely on Shahar Mendelson's "A few notes on Statistical Learning Theory" [1]. Students are encouraged to read this paper.

Let $\mathcal{F}$ be a class of functions. Then $(Z_i)_{i\in\mathcal{I}}$ is a random process indexed by $\mathcal{F}$ if $Z_i(f)$ is a random variable $\forall i$.

As before, μ is a probability measure on Ω, and data $x_1, ..., x_n \sim \mu$. Then μ_n is the empirical measure supported on $x_1, ..., x_n$: $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$. Define $Z_i(\cdot) = (\delta_{x_i} - \mu)(\cdot)$, i.e. $Z_i(f) = f(x_i) - \mathbb{E}_\mu(f)$. Then $Z_1, ..., Z_n$ i.i.d. process with 0 mean.

In the previous lectures we looked at the quantity

$$\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n f(x_i) - \mathbb{E}f\right|. \tag{1}$$

Note that this can be written as $n\sup_{f\in\mathcal{F}} |\sum_{i=1}^n Z_i(f)|$.

Recall that the difficulty with (1) is that we do not know μ and therefore cannot calculate $\mathbb{E}f$. We saw that the classical approach of covering $\mathcal{F}$ and using the Union Bound was too loose.

Symmetrization idea: If $\frac{1}{n}\sum_{i=1}^n f(x_i)$ is close to $\mathbb{E}f$ for various data $x_1, ..., x_n$, then $\frac{1}{n}\sum_{i=1}^n f(x_i)$ is close to $\frac{1}{n}\sum_{i=1}^n f(x_i')$, the empirical average on $x_1', ..., x_n'$ (independent copy of $x_1, ..., x_n$). Therefore, if the two empirical avérarages are far from each other, then empirical error is far from expected error.

Now consider the following:

Example: Fix one function f. Let $\epsilon_1, ..., \epsilon_n$ be i.i.d. Rademacher random variables (taking on values 0 or 1 with probability 1/2). Then

$$\begin{aligned}
\mathbb{P}\left[\left|\sum_{i=1}^n (f(x_i) - f(x_i'))\right| \geq t\right] &= \mathbb{P}\left[\left|\sum_{i=1}^n \epsilon_i(f(x_i) - f(x_i'))\right| \geq t\right] \\
&\leq \mathbb{P}\left[\left|\sum_{i=1}^n \epsilon_i f(x_i)\right| \geq t/2\right] + \mathbb{P}\left[\left|\sum_{i=1}^n \epsilon_i f(x_i')\right| \geq t/2\right] \\
&= 2\mathbb{P}\left[\left|\sum_{i=1}^n \epsilon_i f(x_i)\right| \geq t/2\right]
\end{aligned}$$

Together with the Symmetrization idea, this suggests that controlling $\mathbb{P}\left[\left|\sum_{i=1}^n \epsilon_i f(x_i)\right| \geq t/2\right]$ is enough to control $\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^n f(x_i) - \mathbb{E}f\right| \geq t\right]$.

Empirical Process:

$$Z(x_1, ..., x_n) = \sup_{f \in \mathcal{F}} \left[\mathbb{E}f - \frac{1}{n}\sum_{i=1}^n f(x_i)\right].$$

Rademacher Process:

$$R(x_1, ..., x_n, \epsilon_1, ..., \epsilon_n) = \sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i).$$

$$\begin{aligned}
\mathbb{E}Z &= \mathbb{E}_x \sup_{f \in \mathcal{F}} \left[\mathbb{E}f - \frac{1}{n}\sum_{i=1}^n f(x_i)\right] \\
&= \mathbb{E}_x \sup_{f \in \mathcal{F}} \left[\mathbb{E}_{x'}\left(\frac{1}{n}\sum_{i=1}^n f(x_i')\right) - \frac{1}{n}\sum_{i=1}^n f(x_i)\right] \\
&\leq \mathbb{E}_{x,x'} \sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^n (f(x_i') - f(x_i)) \\
&= \mathbb{E}_{x,x',\epsilon} \sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i(f(x_i') - f(x_i)) \\
&\leq \mathbb{E}_{x,x',\epsilon} \sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i') + \mathbb{E}_{x,x',\epsilon} \sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^n -\epsilon_i f(x_i) \\
&= 2\mathbb{E}R
\end{aligned}$$

As we discussed previously, we would like to bound Z. This will imply "generalization" for any function in $\mathcal{F}$. The above calculation suggests the following: To control Z, show 1) Z is concentrated around its mean $\mathbb{E}Z$, 2) use the above bound $\mathbb{E}Z \leq \mathbb{E}R$, 3) bound $\mathbb{E}R$. (additionally, can show concentration of R around $\mathbb{E}R$ and use R as a data-dependent bound). $\mathbb{E}R$ is called a *Rademacher Average*.

An example of 1): Use McDiarmid's inequality to show concentration of

Z around $\mathbb{E}Z$. Assume $a \leq f(x) \leq b$ for all x and $f \in \mathcal{F}$. Then

$$|Z(x_1, ..., x_i', ..., x_n) - Z(x_1, ..., x_i, ..., x_n)| =$$
$$\left| \sup_{f \in \mathcal{F}} |\mathbb{E}f - \frac{1}{n} \sum_{j=1}^{n} f(x_j) + \left(\frac{1}{n} f(x_i) - \frac{1}{n} f(x_i') \right) | - \sup_{f \in \mathcal{F}} |\mathbb{E}f - \frac{1}{n} \sum_{j=1}^{n} f(x_j)| \right| \leq$$
$$\sup_{f \in \mathcal{F}} \frac{1}{n} |f(x_i) - f(x_i')| \leq \frac{b-a}{n} = c_i$$

McDiarmid's inequality then implies that

$$\mathbb{P}\left(Z - \mathbb{E}Z > t\right) \leq \exp\left(\frac{-t^2}{2 \sum_{i=1}^{n} \frac{(b-a)^2}{n^2}} \right) = \exp\left(\frac{-nt^2}{2(b-a)^2} \right)$$

Equivalently, with probability at least $1 - e^{-u}$,

$$Z - \mathbb{E}Z < \frac{1}{\sqrt{n}} \sqrt{2u}(b-a).$$

So, as the number of samples, n, grows, Z becomes more and more concentrated around $\mathbb{E}Z$. Using the symmetrization step,

$$Z \leq \mathbb{E}Z + \frac{1}{\sqrt{n}} \sqrt{2u}(b-a) \leq 2\mathbb{E}R + \frac{1}{\sqrt{n}} \sqrt{2u}(b-a)$$

with probability at least $1 - e^{-u}$. For sharper inequality, see Talagrand's famous inequality for the suprema of empirical processes.

Why is it easier to bound $\mathbb{E}R$ than $\mathbb{E}Z$? It turns out that $\mathbb{E}R$ has some nice properties (see [1] for more details):

Let $\mathcal{F}$, $\mathcal{G}$ be classes of real-valued functions. Then for any n,

1. If $\mathcal{F} \subseteq \mathcal{G}$, then $\mathbb{E}R(\mathcal{F}) \leq \mathbb{E}R(G)$
2. $\mathbb{E}R(\mathcal{F}) = \mathbb{E}R(conv\mathcal{F})$
3. $\forall c \in \mathbb{R}$, $\mathbb{E}R(c\mathcal{F}) = |c|\mathbb{E}R(\mathcal{F})$
4. If $\phi : \mathbb{R} \to \mathbb{R}$ is L-Lipschitz and $\phi(0) = 0$, then $\mathbb{E}R(\phi(\mathcal{F})) \leq 2L\mathbb{E}R(\mathcal{F})$

5. For RKHS balls, $c(\sum_{i=1}^{\infty} \lambda_i)^{1/2} \le \mathbb{E}R(\mathcal{F}_k) \le C(\sum_{i=1}^{\infty} \lambda_i)^{1/2}$, where λ_i's are eigenvalues of the corresponding linear operator in the RKHS.

Entropy bounds for Rademacher Averages:

$$\mathbb{E}_\epsilon R \le c \frac{1}{\sqrt{n}} \int_0^D \sqrt{\log \mathcal{N}(\epsilon, \mathcal{F}, L_2(\mu_n))} d\epsilon,$$

where $\mathcal{N}$ denotes the covering number, as defined in the previous lectures. The above integral is called the *Dudley integral*.

Example: Let $\mathcal{F}$ be a class with finite VC-dimension V. Then $\mathcal{N}(\epsilon, \mathcal{F}, L_2(\mu_n)) \le \left(\frac{2}{\epsilon}\right)^{kV}$ for some constant k. The Dudley integral above is bounded as

$$\begin{aligned}\int_0^1 \sqrt{\log \mathcal{N}(\epsilon, \mathcal{F}, L_2(\mu_n))} d\epsilon &\le \int_0^1 \sqrt{kV \log 2/\epsilon} d\epsilon \\ &\le k'\sqrt{V} \int_0^1 \sqrt{\log 2/\epsilon} d\epsilon \le k\sqrt{V}.\end{aligned}$$

Therefore, $\mathbb{E}_\epsilon R \le k\sqrt{\frac{V}{n}}$ for some constant k.

References

[1] S. Mendelson *A few notes on Statistical Learning Theory.* Advanced Lectures in Machine Learning, (S. Mendelson, A.J. Smola Eds), LNCS 2600, 1-40. Springer, 2003.